ON THE COMPREHENSION OF EXTREMELY FAST SYNTHETIC SPEECH

Jürgen Trouvain, Phonetics, Saarland University,
and Phonetik-Büro Trouvain, Saarbrücken

We report on a pilot study testing the subjective comprehension of tempo-scaled synthetic speech with 9 sighted and 2 blind students. German texts (length, 100 words) were generated with a formant synthesizer and a diphone synthesizer at seven different tempo steps from 3.5 syllables per second (s/s) to 17.5 s/s. The results show that the blind subjects can understand formant synthesis at all offered rates, whereas the performance of their sighted peers declines at a rate of 10.5 s/s. Contrary to our expectations, diphone synthesis is less easy to understand than formant synthesis for both groups at rates faster than 7.5 s/s. The potential reasons for these two main findings are discussed.

KEYWORDS: speech rate, speech perception, compressed speech, speech synthesis

1 INTRODUCTION[1]

Blind people who use synthetic speech each day often prefer such a fast tempo of speech output that is completely unintelligent for non-daily users of synthetic speech. Audio-file 1 gives an example of a 100-word German text.

The first aim of this study is to quantify the preferred tempo of the blind users in syllables per second (s/s). This quantification is done by comparing the preferred listening tempo with three different reading tempos of humans: 'usual', 'fast' and 'as fast as possible'.

The second aim is to find out at which tempo the comprehension declines for normally sighted non-daily users and blind daily users, respectively.

The third aim is to compare two synthesis methods at various tempos: 1) formant synthesis, which is considered less natural and more robot-like than other forms of synthesis; 2) diphone synthesis which sounds more natural than formant synthesis.

There is anecdotal evidence that blind persons ask for a compression of speech up to factor 8. An overview of various studies for different languages show that human speech is normally articulated between 2 and 7 syllables per second (henceforth abbreviated s/s) depending on factors such as language, speech mode and individual

[1] Acknowledgements: Thanks to the students in the seminar Sprachsynthese aus Benutzersicht at Saarland University in the winter semester 2005/06: Anja Moos, Alexander Tanchev, Fabian Brackhane, Yvonne Flory, Annette Klinger, Dominik Bauer, Steven Webber.

speaker characteristics (cf. Trouvain 2004). If we assume a rather low articulation rate of 3 s/s, a compression by factor 8 would yield in 24 s/s. The average syllable duration would then be reduced from 333 ms to 42 ms.

2 EXPERIMENT

The aim of the experiment is to test the degree of comprehension of synthetic speech at different tempo levels ranging from the preferred tempo of blind daily users to the normal rate of reading aloud.

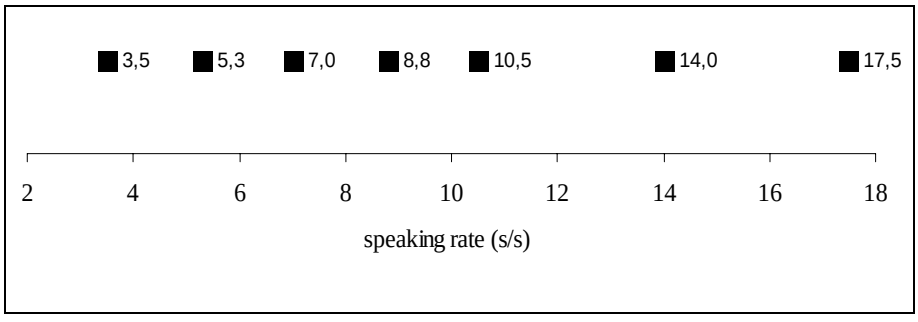
Two methods of speech synthesis were applied: formant synthesis and diphone synthesis. Two different text sorts were used: e-mails (and weblogs) and news.

2.1 TEMPO STEPS

First, the extreme points of the tempo range were defined. The fast end is the usual listening rate for the blind daily user of synthesis at 17.5 s/s. At the slow end we assume 3.5 s/s as a typical reading rate (including pauses). Another orientation point is the rate of 7.0 s/s, which is a rather fast speaking rate which can be found for short stretches in spontaneous discourse.

A further point of orientation is the most extreme reading rate of humans. In an informal fast speaking contest, an average speaking rate of 10.5 s/s was found. The task was to speak a well-known (long) sentence (22 phonological syllables), which is spoken fast in radio and TV advertisements, as fast as possible[2]. The results for the extremely fast sentence spoken by seven native speakers ranged between 9.0 s/s and 12.0 s/s. The mean of 10.5 s/s was taken as the fourth orientation value.

Figure 1: The seven tempo steps presented as speaking rates (including pauses) in syllables per second (s/s).



[2] The text is "Zu Risiken und Nebenwirkungen lesen Sie die Packungsbeilage und fragen Sie Ihren Arzt oder Apotheker." (translation: "For risks and side effects, read the enclosed information sheet and ask your doctor or pharmaceutical chemist.")

Further data points were defined for the test values between the three human speaking rates: 5.3 s/s and 8.8 s/s, and additional value of 14.0 s/s was defined between the fastest human rate and the preferred blind user rate. The seven tempo categories are listed in Fig. 1.

2.2 TEXT MATERIAL

A separate text was used for each tempo category. This was necessary to avoid a learning effect. Two text types were selected: emails and weblogs on the one hand, news texts on the other hand. All texts were about 100 words long (+/- 2 words). In total there were 28 texts: 7 tempo categories x 2 text types x 2 synthesis methods.

2.3 STIMULUS GENERATION

For the two different synthesis methods the following synthesis systems were used. For the formant synthesis we used the German voice of the synthesiser integrated in the screenreader software JAWS (s. Ref). For the diphone synthesis we worked with the synthesis system MARY (Schröder & Trouvain 2003), which makes use of the MBROLA diphone voices (in our case the German voice 'de7') and algorithms (MBROLA).

For both synthesis methods the generated accelerated speech was a LINEAR transformation of the default speech. The fundamental frequency stays constant so that they were no perceivable changes in pitch (no mickey-mouse effect). In contrast to this non-linear compression of speech, accelerated human speech features many NON-LINEAR relationships between speech at a normal tempo and speech at a fast tempo (cf. Trouvain 2004).

In the formant synthesiser it is possible to regulate the tempo of the speech to be generated with the help of a numerical scale. The correlations between the values on that scale with the tempo measured in syllables per second were identified and the stimuli generated accordingly.

For the diphone stimuli the procedure of tempo regulation was different. For each text to be synthesised, the durations of each phone (and each pause) was calculated within the MARY prosody module, and the calculation of the tempo in s/s of this default version was done by hand. Then each phone duration was multiplied with a constant factor to get the speeded-up version at the required tempo. In the last step the diphone speech was generated with the new duration scheme.

Table 1: The phrase "Vorsprung durch Technik" as an example for speeding up the diphone synthesis by factor 2. The 'Mbrola values' are the phones (in machine readable phonetic symbols), duration in ms for each phone in the normal version and the compressed version. The pitch values are kept constant (not included here).

phones	duration (normal version)	duration (version compressed by factor 2)
f	110	55

phones	duration (normal version)	duration (version compressed by factor 2)
o:	105	53
6	39	20
S	72	36
p	62	31
R	39	20
U	63	32
N	49	25
d	44	22
U	63	32
6	35	18
C	56	28
t	69	35
E	77	39
C	66	33
n	40	20
I	103	52
k	95	48

2.4 LISTENING TEST

Subjective comprehension was tested. For the practical reason of test duration, no objective testing of the comprehension was performed. After listening to each text once, subjects had to answer on a six-point scale ranging from "understood everything" (grade 1) to "understood nothing at all" (grade 6) (see table 2).

Table 2: Six degrees of subjective comprehension.

	I've understood
1	everything.
2	almost everything.
3	more than half.
4	less than half.
5	hardly anything.
6	nothing at all.

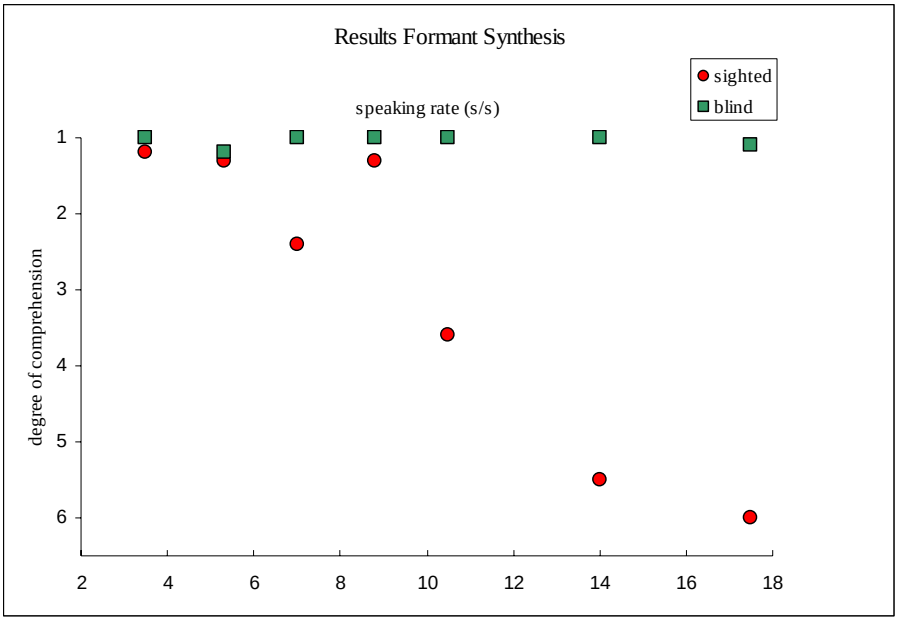
Two groups of subjects were tested:
1) two blind students who use synthetic speech every day. They are very familiar with the voice that has been used in JAWS.
2) nine normally sighted students with no, or hardly any experience with synthetic speech.
All persons are native speakers of German. None of the subjects reported any hearing deficiencies.

For the listening test a PC with its loudspeakers was used in a quiet office environment. Two warm-up stimuli with synthetic speech at a normal conversation rate were played before the test started. The 28 test-stimuli were offered in randomised order. After listening to a stimulus the subject had to give her/his answer by clicking one of six numbered boxes on a computer screen (sighted persons) or orally to the experimenter, who wrote down the answers (blind).

3 RESULTS

In Fig. 2 and 3 the results for the subjective comprehension for both groups of subjects are illustrated. The results for both text types were taken together because there were only a few very small differences in the judgements of each subject.

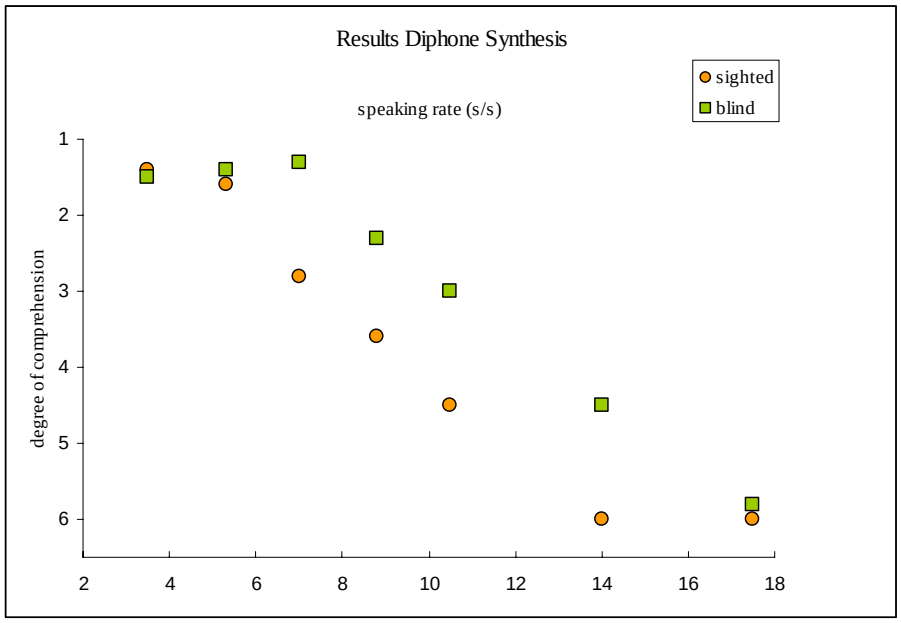
Figure 2: The degree of subjective comprehension (y-axis) as a function of the seven different steps of tempo (x-axis) for FORMANT synthesis.



The results in Fig. 2 clearly show that the two blind subjects have no problems understanding the texts at all rates offered. In contrast, the normal listeners understand only half of the synthesised texts at 10.5 s/s and almost nothing, or nothing at all, at rates faster than 10.5 s/s.

The outlier at 7.0 s/s for the normally sighted persons can be explained with the reading mode in the screenreader. The user can select between "read with (1) no punctuation signs pronounced, or (2) few, or (3) all punctuation signs". We selected "read with few punctuation signs pronounced" because this reading mode is the usual setting for our blind subjects. Unfortunately, the text used for the tempo of 7 s/s for the formant synthesis contained a large number of hyphens which all were read aloud as "Bindestrich". The effect for the sighted persons was a considerable loss of comprehension, whereas the blind subjects had no problems with this type of reading mode.

Figure 3: The degree of subjective comprehension (y-axis) as a function of the seven different steps of tempo (x-axis) for DIPHONE synthesis.



The results show that, in general, fast diphone synthesis (Fig. 3) is not as understandable as fast formant synthesis (Fig. 2). Both user groups judge diphone synthesis to be worse than formant synthesis between 8.8 s/s and 14.0 s/s. At the highest rate tested here (17.5 s/s) diphone speech is unintelligible for both blind subjects as well as for the normally sighted subjects.

There is a difference between the comprehension scores between both groups. Again, the blind subjects understand more at faster rates than the normal seeing

persons. But in contrast to the results of the formant synthetic stimuli the slope proceeds in a parallel way.

Before the test we assumed that diphone synthesis would be preferred over formant synthesis because it sounds more natural (at usual rates of speech). Thus, it could help to the understanding also at faster rates. This assumption must clearly be rejected.

4 DISCUSSION

The comprehension for both blind users was still very good at a syllabic rate of 17.5 s/s. However, this is not the preferred rate all the time when using synthetic voices. Both subjects report that listening at this tempo makes them exhausted after approximately 30 minutes. A slightly slower listening rate is the usual tempo. Of, course, we can not generalise for other blind persons on the basis of just two individuals.

One explanation for the fast listening capability described in this study is the effect of intense and long-term training. Both subjects have worked with exactly this synthetic voice on a daily basis for several years. Another explanation for this extraordinary speech perception skill can be that blind persons use their neural capacities in a different way than sighted persons, who use a lot of capacity for visual processing. Future studies must explore how the learning effect and the neurological conditions can be used to explain how flexible humans are when listening to (this kind of) speech.

It is also astonishing that the fastest rate used here (17.5 s/s) – at which non-blind persons unfamiliar with synthetic voices understand nothing – is not yet the speed limit of (subjective) comprehension for a blind person. Further studies must find out where this limit lies. This kind of research would have important consequences for designing speech synthesis systems for the blind.

We observed a rapid decline in the degree of subjective comprehension between 8.0 s/s und 10.5 s/s for normally sighted persons. We think that it is not a coincidence that this region of deterioration lies at the extreme end of fast human speech production. In other words, humans are never normally exposed to speech faster than 10 s/s, and are therefore not trained to decode such speech in their perception. With extremely fast speech we observe a clear and a close relation between speech production and perception.

Diphone synthesis usually sounds more natural than formant synthesis – at least at normal rates. Consequently we assumed that diphone synthesis would perform better than formant synthesis at faster rates. This assumption was strengthened by the observation that the formant synthesiser used performed worse than the diphone synthesiser with respect to:

- pausing and phrasing
- linguistic pre-processing (e.g. unknown words, abbreviations)
- intonation.

However, for both groups, for non-daily as well as for the daily users of synthetic speech, it was shown that diphone synthesis is felt to be less intelligible than formant synthesis. The reasons for this unexpected result might be found in the unnatural

compression of the diphone speech. In addition, the many concatenation points which characterises diphone synthesis lead to a highly unusual, dense clustering of the artefacts in a shorter period – at normal rates they are much further apart.

It would be interesting to see how fast speech, generated by a non-uniform unit selection synthesiser scores in comparison to diphone and formant synthesis, respectively.

Also, a test using stimuli with extremely fast natural speech is of interest. In natural speech we show more reductions and hypo-articulations which contributes to naturalness – at normal rates. However, Janse (2003) found that very fast speech with hypo-speech is less intelligible than hyper-speech.

Usually, comprehension and naturalness are the two main criteria for assessing the quality of synthetic speech output. But for the some groups of users, naturalness is clearly NOT desirable, naturalness is a hindrance rather than a help. We see with this example that the benefit of speech synthesis strongly depends on the needs of its users. However, special needs of special users are often ignored in the evaluation of speech synthesis systems. To find out more about the actual needs and problems of real listeners of synthetic speech is an important task for phoneticians working in the area of speech technology. Phonetics can contribute a great deal to the improvement of speech-signal modification methods as well as to the understanding of the perception of ultra-fast speech.

REFERENCES

Janse, Esther. 2003. *Production and perception of fast speech* (LOT Dissertation Series 69). Utrecht: Netherlands Graduate School of Linguistics dissertation. http://www.let.uu.nl/~Esther.Janse/personal/. (17 May, 2007.)

JAWS (Job Access With Speech) Screenreader software. http://www.freedomsci.de. (17 May, 2007.)

MBROLA Multilingual Speech Synthesizer. http://tcts.fpms.ac.be/synthesis/mbrola.html. (17 May, 2007.)

Schröder, Marc & Jürgen Trouvain. 2003. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology* 6. 365-377. www.dfki.de/~schroed/articles/schroeder_trouvain2003.pdf (17 May, 2007.)

Trouvain, Jürgen. 2004. *Tempo variation in speech production: Implications for speech synthesis* (Phonus 8). Saarbrücken: Saarland University dissertation. http://www.coli.uni-saarland.de/groups/WB/Phonetics/structure.php?page=Research/PHONUS_research_reports/phonus.php (17 May, 2007.)

J. Trouvain. 2007. Comprehension of synthetic speech. *Saarland Working Papers in Linguistics (SWPL)* 1. 5-13.

Dr. Jürgen Trouvain
Saarland University
FR. 4.7: Phonetics
Building C7 2
Im Stadtwald
D-66123 Saarbrücken

trouvain@coli.uni-sb.de